Extending the user capacity of MU-MIMO systems with low detection complexity and receive diversity

Article (Accepted Version)

Extending the User Capacity of MU-MIMO Systems with Low Detection Complexity and Receive Diversity

Walid A. Al-Hussaibi
Department of Electrical Techniques, BTI,
Southern Technical University, Basrah, 42001, Iraq
E-mail: alhussaibi@stu.edu.iq

Falah H. Ali
Communications Research Group, School of Engineering and Informatics,
University of Sussex, Brighton, BN19QT, UK
E-mail: f.h.ali@sussex.ac.uk

Abstract

Multiple-input multiple-output (MIMO) based technologies are considered as an integral part of the upcoming 5G communications to fulfil the ever-increasing demands of wireless applications with high spectral efficiency requirements. However, in uplink multiuser MIMO (MU-MIMO) channels, the number of allowed users is limited by the number of receive antennas associated with radio frequency (RF) chains at the base-station and the complexity burden of multiuser detection (MUD). In this paper, a novel group layer MU-MIMO scheme with low complexity MUD is proposed to increase the number of served users well beyond the available RF chains. By taking the advantage of power control and inherent path loss in cellular systems, the allowed users are divided into groups based on their received power. Efficient group power allocation and group layer MUD (GL-MUD) are utilized to provide a valuable tradeoff between complexity and achieved performance. Furthermore, when more receive antennas than RF chains is implemented, a generalized norm based antenna selection algorithm is proposed to enhance the error performance. Symbol error probability expressions are derived and the effectiveness of proposed scheme is demonstrated through numerical simulations compared with the conventional MU-MIMO and non-orthogonal multiple-access (NOMA) systems over Rayleigh fading channels. The results show a substantial increase in user capacity up to two-fold for the available number of RF chains. In addition, significant signal-to-noise ratio gain is achieved using GL-MUD compared with different MUD techniques.

Keywords: MU-MIMO; user overloading; capacity; error performance; multiuser detection; antenna selection.

1. Introduction

Wireless and mobile communication systems are rapidly expanding worldwide and becoming an essential technology with crucial impact on modern life. Therefore, efficient spectrum utilization is required to meet the increasing number of clients and their demands for wireless services [1, 2]. Furthermore, power and complexity constraints have added more challenges on the development of future systems. However, it is widely acknowledged that spatial multiplexing multiple-input multiple-output (MIMO) represents a key technology for higher data rate without consuming extra bandwidth and transmit power. Therefore, it is considered as a fundamental enabling technology to fulfil the high spectral efficiency demand of next fifth generation (5G) systems towards Gigabits communications [2-5].

1.1. Background

In cellular systems, multiuser MIMO (MU-MIMO) enable multiple users equipped with one antenna or more to access the base-station (BS) simultaneously without subdivision in the scarce resources of time, frequency, or codes [4-7]. The IEEE Worldwide Interoperability for Microwave Access (WiMAX) and Third Generation Partnership Project (3GPP) standards of forth generation (4G) systems represented by IEEE 802.16m and Long Term Evolution Advanced (LTE-A), respectively are examples of such schemes [1, 8]. However, the maximum number of allowed users (K) is limited by the total degree of freedom (DoF) represented by the number of BS antennas (m) associated with radio frequency (RF) chains [9-11]. Also, the multiuser detection (MUD) method employed at BS receiver has direct impact on the maximum number of supported users. Therefore, different scheduling methods are utilized to exploit the inherent multiuser diversity when the sum of users' antennas is larger than m [12].

For linear MUD at BS, the number of single antenna users that can be served reliably is $K \leq m$ [9]. However, it cannot be used in overloading or *rank-deficient* scenarios (in which $K > m$) as the overall MU-MIMO channel becomes noninvertible, and hence reducing the required DoF for signal detection. Linear MUD methods such as zero forcing (ZF), minimum mean squared error (MMSE), and minimum bit error rate (MBER) have low implementation complexity but only capable of providing limited performance [3, 13]. In contrast, nonlinear MUD such as maximum likelihood (ML) can increase the user capacity beyond the number of BS antennas, however at the cost of higher computational complexity which increases exponentially with K [14]. Practical implementation of ML detector in overloaded systems is prohibitive, and therefore many suboptimal techniques have been developed such as successive interference cancellation (SIC) [15], sphere decoding (SD) [16], and iterative groupwise detection [17]. Although complexity of suboptimal methods is less than the optimal ML, they are still more complex than linear MUDs causing significant limitation on the essential DoF required for high capacity systems.

To increase the DoF represented by m BS antennas, same number of m costly RF chains is required for processing and down-conversion. However, implementing more RF chains to support the additional users is impractical in terms of hardware requirements, consumed power, and size [4]. In the last years, feasible antenna selection diversity techniques have been proposed to capture most of the gains promised by multi-antenna systems when the number of available RF chains is smaller than the number of antenna elements [2, 18-21]. And therefore since extra antenna elements, RF switches, and digital signal processing circuitry are usually inexpensive, the gain of antenna selection can be achieved with only small additional cost [19]. This technology has been adopted in IEEE 802.16e/n/m WiMAX and 3GPP-LTE standards [21] and becomes an essential part of the promising massive MIMO systems [20].

On different directions, user overloading based on users grouping to share the same DoF has been investigated in [22-27] for different multiple-access schemes. In [22], non-orthogonal multiple-access (NOMA) is proposed to allow simultaneous transmission of more than one user for each subcarrier in orthogonal frequency division multiplexing (OFDM) system with efficient interference cancellation technique. In [23], more users than the spreading factors of orthogonal code division multiple-access (OCDMA) are accommodated though at the cost of higher signal-to-noise ratio (SNR). User capacity of OCDMA has been improved also by using other approaches such as superposition coding [24], multiple antennas with linear MUD [25], and collaborative spreading [26]. In [27], spectrum based orthogonal user partitioning is considered to reduce the overhead requirements in MIMO channels and increase the sum rate. User pairing based on the power domain is also applied in [11] to downlink MIMO channels with NOMA for improved sum-rate performance. However, full-rank channel matrices are assumed which represent a non-challenging scenario.

1.2. Main Objectives and Contributions

For more efficient spectrum utilization, new design approaches for MU-MIMO applications are of high interest. This work aims to improve the user capacity with affordable complexity and enhanced reliability.

In this paper, uplink group layer MU-MIMO (GL-MU-MIMO) scheme is proposed by exploiting the spatial difference among users and employing low complexity MUD with receive antenna selection (RAS) facility. By taking the advantage of power control at BS and the inherent path loss in cellular systems due to users' geographical locations, the active users are divided into two groups, namely high power group (HPG) and low power group (LPG). The assigned powers are efficiently controlled through group power allocation ratio (η) to achieve extended user capacity and error performance. At the BS receiver, group layer MUD (GL-MUD) is utilized. For RAS diversity when more receive antennas than available RF chains are implemented, a generalized norm based selection (GNBS) algorithm is proposed to select the best subset of receive antennas in terms of their channel gains. The superiority of proposed scheme is validated through numerical simulations over Rayleigh

fading channels and compared with the MU-MIMO employing linear MUDs and NOMA with ML receiver. The proposed system is motivated by 5G cellular mobile requirements of high spectral efficiency and future applications including the massive increase in connected devices. It enables efficient spectrum utilization and reliable communications for different multiantenna applications for 5G networks such as opportunistic communications, cognitive radio, cooperative transmission, and wireless power transfer.

The contributions of this paper are summarised as follows:

1) A novel GL-MU-MIMO scheme is proposed to extend the user capacity well beyond the limit of conventional MU-MIMO with linear MUDs. It is shown that up to two-fold increase in the number of allowed users is achievable for same number of essential RF chains at BS.
2) A low complexity GL-MUD using MMSE based group MUD and group SIC (GSIC) is employed rather than linear MMSE-SIC, MMSE, or ZF receivers which require more RF chains for same number of users K. In addition, RAS is integrated to enhance the error performance considerably. To the best of our knowledge, combining different strategies as used in this work to address the highlighted challenges is the first of its kind and required critical system design and analysis.
3) Symbol error probability equations are derived to evaluate the system performance compared with the existing MU-MIMO and NOMA systems.

The remainder of this paper is organized as follows: In Section 2, system design of GL-MU-MIMO is described including the system model, GL-MUD, and RAS diversity. In Section 3, performance analysis of the proposed system is presented including the error probability, user capacity, and complexity analysis of GL-MUD. The conducted results are shown in Section 4. Finally, Section 5 concludes the paper.

Notations: Bold-face uppercase and lowercase letters denote matrices and vectors, respectively. Plain lowercase letters stand for scalars. $\mathcal{C}^{m\times u}$ denotes complex $m \times u$ matrix while $\mathcal{R}^{m\times u}$ is for real $m \times u$ matrix. Superscripts $[.]^{\mathrm{H}}$, $[.]^{\mathrm{T}}$ and $[.]^{\dagger}$ stand for conjugate transposition, transposition and pseudoinverse, respectively. $\mathbf{I}_m$ is $m \times m$ identity matrix and $\|.\|$ stands for the Euclidean vector norm. $|.|$ denote the determinant for matrices and magnitude for vectors. $Q(.)$ denotes the Q-function.

2. System Design of GL-MU-MIMO

2.1. System Model

Consider an overloaded uplink MU-MIMO of K active users communicating simultaneously over Rayleigh fading channel with one common BS in one cell cellular system as depicted in Fig. 1. The users' geographical locations are randomly distributed within the considered system cell where each mobile terminal has a single antenna while the BS is equipped with m antennas (larger than m_s RF chains) and employs RAS to select the best subset of antennas m_s based on their channel conditions.

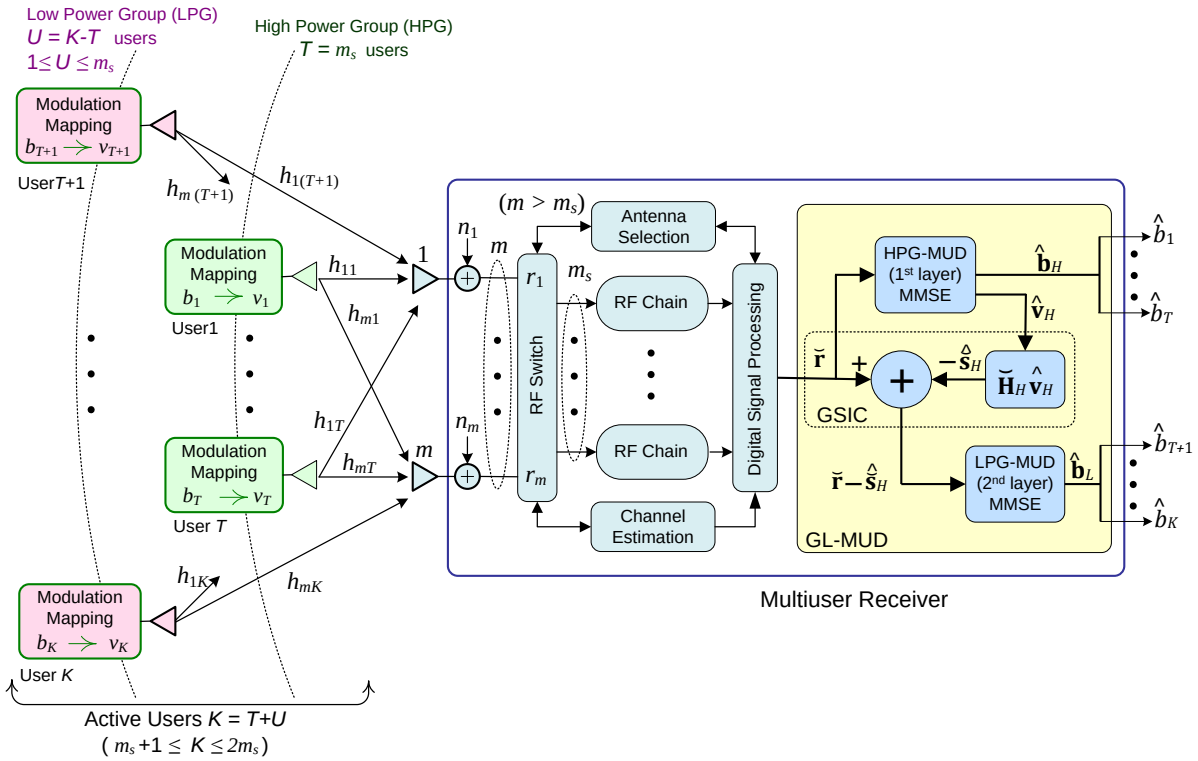

Figure 1. GL-MU-MIMO system with K active users and one common BS with RAS and GL-MUD.

In the context of spectrally efficient wireless systems, mobile users of high channel gains have the priority of the accessible communication links in contrast to those users of poorer channel conditions. On the other hand, a balance between spectral efficiency and fairness in distributing the system resources among active users should be maintained for future communications [11, 22]. Motivated by these facts, user partitioning is considered in this work by dividing the allowed users into two groups based on their relative proximity from BS represented by the path loss and channel gain conditions as: HPG of strong channel users and LPG of weak channel users. Practically, user grouping can be achieved based on the average signal attenuation for each user $\mathcal{L}_k; k = 1, \dots, K$ which is inversely proportional to the received power. These attenuation factors are slowly varying and can be measured at the BS for example during the training phase [10].

At the BS receiver, GL-MUD is proposed using two layers of MMSE based MUD and linked by GSIC technique. The first detection layer is referred to as HPG-MUD while the second is denoted as LPG-MUD. Since the total DoF of the linear MUD for each of the designed groups is limited by the available m_s RF chains, the total allowed users (streams) is limited by $K \leq 2m_s$. Consequently, design criterion for group formation is performed by sorting the active users at first in ascending order according to $\mathcal{L}_k$. Then, to satisfy the channel rank condition, HPG is configured from the first $T = m_s$ users of highest received powers (e.g. users near the BS). In practice, this group will have the impact to maintain the maximum user capacity and spectral efficiency as those achieved by the generic MU-MIMO with linear MUD. On the other hand, LPG is formed from the rest users $U = (K - T) \leq m_s$

of lowest powers (e.g. users near the cell edge) to satisfy the fairness among active users. Moreover, size of LPG is designed as $1 \leq U \leq m_s$ for the purpose of signal detection and interference management. Note that the additional users U of weak channel conditions are commonly terminated (not scheduled) in the conventional system. For GL-MU-MIMO and based on the capacity of HPG and LPG, the range of supported users is given as $m_s + 1 \leq K \leq 2m_s$.

As in superposition coding [12, 28], the received power difference between HPG and LPG is essential to simplify the decoder task and manage the interference level between designed groups. Therefore, received powers at BS from HPG users P_H and LPG users P_L are given under total average power constraint of $P = P_H + P_L$ during every symbol period as

$$P_H = (1 - \eta)P \tag{1}$$

$$P_L = \eta P = \eta P_H/(1 - \eta) \tag{2}$$

$$\eta = P_L/(P_H + P_L) \tag{3}$$

where $\eta;\ 0 < \eta < 0.5$ is *group power allocation ratio* maintained by power control at the BS based on the acceptable interference level between designed groups and target error performance. It should be noted that the value of power constraint P depends mainly on the maximum transmit power that the user terminals can handle according to the specifications of their power amplifiers and/or the spectrum regulations allow. Furthermore, statistics-aware transmit power allocation [7] is assumed for users within each group to compensate the path loss and satisfy the average received power conditions for HPG and LPG as $\{p_k = (1 - \eta)P/(T\mathcal{L}_k)\}_{k=1}^{T}$ and $\{p_k = \eta P/(U\mathcal{L}_k)\}_{k=T+1}^{K}$, respectively. This strategy has the advantage of allowing uniform user performance within each group due to equal average effective channel gain for all mobile terminals $(p_k\mathcal{L}_k)$. To be achieved, only small feedback overhead is required to provide the users with their transmit powers $p_k; k = 1, \ldots, K$ which is comparable to that of existing MU-MIMO approaches [5].

The received signal vector from HPG and LPG users at m receive antennas is represented as

$$\mathbf{r} = \sum_{k=1}^{T} \mathbf{h}_k v_k + \sum_{k=T+1}^{K} \mathbf{h}_k v_k + \mathbf{n} = \mathbf{s}_H + \mathbf{s}_L + \mathbf{n} \tag{4}$$

where $\mathbf{r} = [r_1 \quad \cdots \quad r_m]^{\mathrm{T}} \in \mathcal{C}^{m\times 1}$, $\mathbf{h}_k = [h_{1k} \quad \cdots \quad h_{mk}]^{\mathrm{T}} \in \mathcal{C}^{m\times 1}$ is the channel vector of user k whose entries h_{lk} are zero mean unit variance complex fading coefficient between user k and l^{th} receive antenna, v_k is transmitted signal of user k subject to power constraint p_k and modulated from equiprobable data b_k using *M*-PSK or *M*-QAM constellations, $\mathbf{n} = [n_1 \quad \cdots \quad n_m]^{\mathrm{T}} \in \mathcal{C}^{m\times 1}$ is i.i.d complex additive white Gaussian noise (AWGN) vector with elements having zero mean and variance

σ_n^2, $\mathbf{s}_H \in \mathcal{C}^{m\times 1}$ and $\mathbf{s}_L \in \mathcal{C}^{m\times 1}$ are superimposed signal vectors of HPG users and LPG users over their entire channels, respectively.

The overall channel matrix $\mathbf{H} \in \mathcal{C}^{m\times K}$ can be represented as

$$\mathbf{H} = [\mathbf{h}_1 \ \cdots \ \mathbf{h}_k \ \cdots \ \mathbf{h}_K] = \left[\underline{\mathbf{h}}_1, \cdots, \underline{\mathbf{h}}_l, \cdots, \underline{\mathbf{h}}_m\right]^{\mathrm{T}} \tag{5}$$

where $\underline{\mathbf{h}}_l \in \mathcal{C}^{1\times K}$ is the l^{th} row of $\mathbf{H}$ corresponding to l^{th} receive antenna. Also, $\mathbf{H}$ can be represented in terms of HPG channel $\mathbf{H}_H \in \mathcal{C}^{m\times T}$ and LPG channel $\mathbf{H}_L \in \mathcal{C}^{m\times U}$ as $\mathbf{H} = [\mathbf{H}_H \ \mathbf{H}_L]$. Therefore, (4) can be rewritten as

$$\mathbf{r} = \mathbf{H}_H\mathbf{v}_H + \mathbf{H}_L\mathbf{v}_L + \mathbf{n} \tag{6}$$

where the transmitted signal vectors $\mathbf{v}_H = [v_1 \ \cdots \ v_T]^{\mathrm{T}} \in \mathcal{C}^{T\times 1}$ and $\mathbf{v}_L = [v_{T+1} \ \cdots \ v_K]^{\mathrm{T}} \in \mathcal{C}^{U\times 1}$ are belongs to HPG and LPG, respectively.

With RAS diversity, the received signal vector $\breve{\mathbf{r}}$ associated with the selection of m_s from m receive antennas can be written as

$$\breve{\mathbf{r}} = \sum_{k=1}^{T} \breve{\mathbf{h}}_k v_k + \sum_{k=T+1}^{K} \breve{\mathbf{h}}_k v_k + \breve{\mathbf{n}} = \breve{\mathbf{H}}_H\mathbf{v}_H + \breve{\mathbf{H}}_L\mathbf{v}_L + \breve{\mathbf{n}} = \breve{\mathbf{s}}_H + \breve{\mathbf{s}}_L + \breve{\mathbf{n}} \tag{7}$$

where $\breve{\mathbf{r}} \in \mathcal{C}^{m_s\times 1}$, $\breve{\mathbf{h}}_k \in \mathcal{C}^{m_s\times 1}$, and $\breve{\mathbf{n}} \in \mathcal{C}^{m_s\times 1}$ denote received signal, k^{th} user channel, and noise vectors after selection, respectively. $\breve{\mathbf{H}}_H \in \mathcal{C}^{m_s\times T}$ and $\breve{\mathbf{H}}_L \in \mathcal{C}^{m_s\times U}$ are HPG and LPG channels associated with RAS, respectively. $\breve{\mathbf{s}}_H \in \mathcal{C}^{m_s\times 1}$ and $\breve{\mathbf{s}}_L \in \mathcal{C}^{m_s\times 1}$ are the superimposed signal vectors of HPG and LPG over their entire channels and associated with RAS, respectively.

In this work, the following assumptions are considered:

1) Perfect estimation of channel state information (CSI) is assumed at the BS receiver for the purpose of RAS, calculation of the power allocation ratio (η), HPG interference, and signal detection. Effect of imperfect CSI is beyond the scope of this paper.
2) Channel fading rate is assumed to be much less than the data rate, so it remains constant over a frame of hundreds of symbols and changes from one frame to the next independently [10, 19].
3) Impact of user scheduling is not included by assuming system cell of $K \leq 2m_s$ perfectly synchronized users.

2.2. GL-MUD

Taking the advantage of power disparity between HPG and LPG, GL-MUD is proposed using MMSE based group MUD with GSIC as shown in Fig. 1. The MMSE detector has the functionality of maximizing the output signal-to-interference-plus-noise ratio (SINR) and minimizing the mean square error (MSE) between transmitted and estimated symbols [13]. For a given channel matrix $\mathbf{G} \in \mathcal{C}^{m_s\times K}$

with $m_s = m \geq K$, the MMSE weight matrix $\mathbf{W}_{mmse} \in \mathcal{C}^{K \times m_s}$ is given by $\left(\mathbf{G}^{\mathrm{H}}\mathbf{G} + \sigma_n^2 \mathbf{I}_K\right)^{-1}\mathbf{G}^{\mathrm{H}}$ and can be used for signal detection within considered channel's coherence time. The diversity order of this receiver is reported as $(m_s - K + 1)$ [3].

In the first detection layer of GL-MUD, selected received vector $\breve{\mathbf{r}}$ is processed by HPG-MUD using MMSE technique to estimate the data of HPG users while treating LPG signals as a background noise. Using HPG weight matrix $\mathbf{W}_H = \left(\breve{\mathbf{H}}_H^{\mathrm{H}}\breve{\mathbf{H}}_H + \sigma_n^2 \mathbf{I}_T\right)^{-1}\breve{\mathbf{H}}_H^{\mathrm{H}} \in \mathcal{C}^{T \times m_s}$, the transmitted HPG signal vector is estimated with $(m_s - T + 1)$ diversity order as

$$\hat{\mathbf{v}}_H = \mathbf{W}_H\breve{\mathbf{r}} = \mathbf{v}_H + \underbrace{\mathbf{W}_H(\breve{\mathbf{s}}_L + \breve{\mathbf{n}})} \tag{8}$$

where $\hat{\mathbf{v}}_H = [\hat{v}_1 \quad \cdots \quad \hat{v}_T]^{\mathrm{T}}$ is affected by the minimum MSE from weighted LPG interference and noise term $\mathbf{W}_H(\breve{\mathbf{s}}_L + \breve{\mathbf{n}})$. At the output of HPG-MUD, estimated data vector $\hat{\mathbf{b}}_H = [\hat{b}_1 \quad \cdots \quad \hat{b}_T]^{\mathrm{T}} \in \mathcal{R}^{T \times 1}$ is remapped from $\hat{\mathbf{v}}_H$.

In the second detection layer, estimated signals of HPG will be multiplied by their channel estimates to calculate the group interference as $\hat{\breve{\mathbf{s}}}_H = \breve{\mathbf{H}}_H\hat{\mathbf{v}}_H$. After applying GSIC of $\hat{\breve{\mathbf{s}}}_H$ from $\breve{\mathbf{r}}$, the input signal $\left(\breve{\mathbf{r}} - \hat{\breve{\mathbf{s}}}_H\right)$ to LPG-MUD is processed for LPG data estimation using MMSE of weight matrix $\mathbf{W}_L = \left(\breve{\mathbf{H}}_L^{\mathrm{H}}\breve{\mathbf{H}}_L + \sigma_n^2 \mathbf{I}_U\right)^{-1}\breve{\mathbf{H}}_L^{\mathrm{H}} \in \mathcal{C}^{U \times m_s}$. Transmitted LPG signal vector can be estimated with diversity order of $(m_s - U + 1)$ as

$$\hat{\mathbf{v}}_L = \mathbf{W}_L\left(\breve{\mathbf{r}} - \hat{\breve{\mathbf{s}}}_H\right) = \mathbf{v}_L + \mathbf{W}_L\left(\breve{\mathbf{s}}_H - \hat{\breve{\mathbf{s}}}_H + \breve{\mathbf{n}}\right) = \mathbf{v}_L + \underbrace{\mathbf{W}_L(\mathbf{e} + \breve{\mathbf{n}})} \tag{9}$$

where $\hat{\mathbf{v}}_L = [\hat{v}_{T+1} \quad \cdots \quad \hat{v}_K]^{\mathrm{T}}$ is influenced by the term $\mathbf{W}_L(\mathbf{e} + \breve{\mathbf{n}})$ of weighted HPG interference cancellation error vector $\mathbf{e} = \breve{\mathbf{s}}_H - \hat{\breve{\mathbf{s}}}_H \in \mathcal{C}^{m_s \times 1}$ of zero mean and σ_e^2 variance elements and weighted noise vector. Data of LPG users can be found then by remapping $\hat{\mathbf{v}}_L$ to $\hat{\mathbf{b}}_L = [\hat{b}_{T+1} \quad \cdots \quad \hat{b}_K]^{\mathrm{T}} \in \mathcal{R}^{U \times 1}$.

Highest performance of LPG can be achieved when perfect cancellation of HPG interference is accomplished (i.e. $\sigma_e^2 = 0$) such that the desired signals are disturbed by AWGN only. On the other hand, performance of HPG can be improved if the interference level from LPG is minimized (i.e. low P_L). For this purpose, appropriate choice of η should be used to allow reliable communication for different user overloading (U) with minimum σ_e^2. At the output of GL-MUD, overall estimated data vector can be found as $\hat{\mathbf{b}} = [\hat{b}_1 \quad \cdots \quad \hat{b}_K]^T \in \mathcal{R}^{K \times 1}$.

2.3. RAS Diversity

To improve the bit-error-rate (BER) performance with affordable complexity, RAS diversity is utilized in the proposed system. With the availability of CSI at the receiver, a generalized norm based selection (GNBS) algorithm is designed to select the best subset of m_s from m receive antennas as

$\mathcal{s}_i \in \mathbb{S}$, where $\mathbb{S} = \{\mathcal{s}_1, \dots, \mathcal{s}_i, \dots, \mathcal{s}_{|\mathbb{S}|}\}$ represent all possible subsets with cardinality of $|\mathbb{S}| = \binom{m}{m_s}$. In this method, selection of m_s antennas is based on the corresponding rows of **H** with the largest Euclidean norm (power) to maximize the received SNR. The complexity of this algorithm is of $\mathcal{O}(Km)$ due to the requirement of vector norm calculations of all m rows. Thus, low selection complexity is achieved compared with the most popular methods in [18-20].

GNBS Algorithm:

1) Define the set of receive antennas as, $\mathcal{X} = [1, \dots, m]$ with $l \in \mathcal{X}$ representing the l^{th} antenna.
2) Given $\mathbf{H} = \left[\underline{\mathbf{h}}_1, \dots, \underline{\mathbf{h}}_l, \dots, \underline{\mathbf{h}}_m\right]^{\mathrm{T}}$ where $\underline{\mathbf{h}}_l$ is the l^{th} row corresponding to l^{th} antenna.
3) For all l in $\mathcal{X}$, calculate the power of $\underline{\mathbf{h}}_l$ as $\left\|\underline{\mathbf{h}}_l\right\|^2$.
4) Sort $\mathcal{X}$ elements according to the associated power of each channel vector in descending order.
5) To select the best subset of m_s receive antennas; choose l representing the channel vectors with maximum power from first m_s elements in $\mathcal{X}$.
6) Construct channel matrix associated with the selected antennas as $\breve{\mathbf{H}} = \left[\breve{\mathbf{H}}_H \;\; \breve{\mathbf{H}}_L\right]$.

3. Performance Analysis

3.2. Error Probability

From [29, Eqs. (5-2-5, 5-2-61, and 5-2-81)], the symbol error probability for coherent reception of *M*-PSK and *M*-QAM can be written generally for a certain SNR γ as

$$\mathcal{P}_e(\gamma) = \lambda\, Q\left(\sqrt{\varphi\, \gamma}\right) \tag{10}$$

where λ and φ factors are determined for specific constellations. For example, $\lambda = 1$ and $\varphi = 2$ for BPSK; $\lambda \approx 2$ and $\varphi \approx 2\sin^2(\pi/M)$ for *M*-PSK; $\lambda \approx 4$ and $\varphi \approx 3/(M-1)$ for *M*-QAM.

For the proposed system, the symbol error probability of k^{th} user in HPG conditioned on constant channel realization $\breve{\mathbf{H}} = \left[\breve{\mathbf{H}}_H \;\; \breve{\mathbf{H}}_L\right]$ can be written for certain average post-processing SINR $\gamma_k^{(H)}$ at first layer of GL-MUD as

$$\mathcal{P}_e^{(H)}\left(\hat{v}_k \middle| \breve{\mathbf{H}}\right) = \lambda\, Q\left(\sqrt{\varphi\, \gamma_k^{(H)}}\right); \qquad k = 1, \dots, T \tag{11}$$

and $\gamma_k^{(H)}; k = 1, \dots, T$ is given by

$$\gamma_k^{(H)} = \frac{\frac{(1-\eta)P}{T}\left|(\mathbf{W}_H)_k\, \breve{\mathbf{h}}_k\right|^2}{\sigma_n^2\, \|(\mathbf{W}_H)_k\|^2 + \frac{(1-\eta)P}{T}\boldsymbol{\Omega}_H + \frac{\eta P}{U}\boldsymbol{\Omega}_L} \tag{12}$$

where $(\mathbf{W}_H)_k$ denotes the k^{th} row of $\mathbf{W}_H$ matrix, $\mathbf{\Omega}_H = \sum_{j=1, j\neq k}^{T} \left|(\mathbf{W}_H)_k \, \check{\mathbf{h}}_j\right|^2$ is the interference from other users in HPG, and $\mathbf{\Omega}_L = \sum_{j=T+1}^{K} \left|(\mathbf{W}_H)_k \, \check{\mathbf{h}}_j\right|^2$ is LPG interference. In the right-hand of above equation, the numerator represents the signal power of k^{th} user while the denominator include the noise and interference terms.

For k^{th} user in LPG at second layer of GL-MUD, the symbol error probability conditioned on constant channel realization $\check{\mathbf{H}}$ can be written for average SINR $\gamma_k^{(L)}$ as

$$\mathcal{P}_e^{(L)}\left(\hat{v}_k \middle| \check{\mathbf{H}}\right) = \lambda \, Q\left(\sqrt{\varphi \, \gamma_k^{(L)}}\right); \qquad k = T+1, \dots, K \tag{13}$$

and $\gamma_k^{(L)}; k = T+1, \dots, K$ is given by

$$\gamma_k^{(L)} = \frac{\frac{\eta P}{U} \left|(\mathbf{W}_L)_k \, \check{\mathbf{h}}_k\right|^2}{\sigma_n^2 \, \|(\mathbf{W}_L)_k\|^2 + \frac{\eta P}{U} \mathbf{\Lambda}_L + \sigma_e^2 \, \|(\mathbf{W}_L)_k\|^2} \tag{14}$$

where $(\mathbf{W}_L)_k$ stand for k^{th} row of $\mathbf{W}_L$ matrix and $\mathbf{\Lambda}_L = \sum_{j=T+1, j\neq k}^{K} \left|(\mathbf{W}_L)_k \, \check{\mathbf{h}}_j\right|^2$ is the interference from other LPG users. The numerator of the above equation represents the signal power of k^{th} user while the denominator include the noise, interference of LPG users, and the interference cancellation error from HPG. Thus, the average symbol error probability $\mathcal{P}_e^{(GL)}$ conditioned on $\check{\mathbf{H}}$ can be found in terms of $\mathcal{P}_e^{(H)}$ and $\mathcal{P}_e^{(L)}$ as

$$\mathcal{P}_e^{(GL)} = \frac{1}{K}\left[\sum_{k=1}^{T} \mathcal{P}_e^{(H)}\left(\hat{v}_k \middle| \check{\mathbf{H}}\right) + \sum_{k=T+1}^{K} \mathcal{P}_e^{(L)}\left(\hat{v}_k \middle| \check{\mathbf{H}}\right)\right] \tag{15}$$

On the other hand, average symbol error probability $\mathcal{P}_e^{(Conv)}$ of conventional MU-MIMO over constant channel $\mathbf{G} \in \mathcal{C}^{m_s \times K}$ and with same total average power constraint P as in the proposed scheme can be calculated for average SNR $\gamma_k^{(Conv)}$ as

$$\mathcal{P}_e^{(Conv)} = \frac{\lambda}{K}\left[\sum_{k=1}^{K} Q\left(\sqrt{\varphi \, \gamma_k^{(Conv)}}\right)\right] = \frac{\lambda}{K}\left[\sum_{k=1}^{K} Q\left(\sqrt{\frac{\varphi \, P \, |\mathbf{w}_k \, \mathbf{g}_k|^2}{K\sigma_n^2 \, \|\mathbf{w}_k\|^2 + P\sum_{j=1, j\neq k}^{K} \left|\mathbf{w}_k \, \mathbf{g}_j\right|^2}}\right)\right] \tag{16}$$

where $\mathbf{w}_k$ is the k^{th} row of $\mathbf{W}_{zf} = \left[\left(\mathbf{G}^{\mathrm{H}}\mathbf{G}\right)^{-1}\mathbf{G}^{\mathrm{H}}\right]$ or $\mathbf{W}_{mmse} = \left[\left(\mathbf{G}^{\mathrm{H}}\mathbf{G} + \sigma_n^2\mathbf{I}_K\right)^{-1}\mathbf{G}^{\mathrm{H}}\right]$ for ZF and MMSE, respectively, and $\mathbf{g}_k$ is the k^{th} column of $\mathbf{G}$.

Note that for realistic GL-MU-MIMO system of instantaneous channels $\check{\mathbf{H}}$, the results of equations (11), (13), and (15) should be averaged over large number of channel realizations [19, 20]. Exact

symbol (or bit) error probability can be calculated for coherent BPSK signals ($\lambda = 1$ and $\varphi = 2$) and coherent Gray coded QPSK which is similar to that of BPSK [29, Eq. (5-2-5)]. For higher order modulation and due to approximate values of λ and φ, the aforementioned equations will provide approximate results. Similar considerations is applied for (16) to find the performance of conventional MU-MIMO system.

3.1. User Capacity

For fading multiple-access channel, user capacity C_{user} can be defined as the maximum number of users (K) that simultaneously and reliably communicate with a common BS in a cell using available resources of time, frequency, space, and codes [30 and references therein]. Note that the term "user capacity" is used also in the literature to describe capacity of the user in terms of the maximum achievable rate [26, 28] which is outside the scope of this paper.

In the proposed GL-MU-MIMO, one of the main objectives is to increase the user capacity beyond the spatial DoF limit represented by m_s RF chains at BS. Since HPG is formed from $T = m_s$ users and LPG includes the rest from K as $1 \leq (U = K - T) \leq m_s$, the user capacity can be maximized when the interference of LPG is minimal and perfect cancellation of HPG interference is achieved at the second layer of GL-MUD. Therefore, the upper bound user capacity can be written as

$$C_{user-GL}^{max} = 2m_s. \tag{17}$$

On the other hand, for imperfect HPG interference cancellation and to achieve target error performance, the lower bound user capacity is given as

$$C_{user-GL}^{min} = m_s + 1 \tag{18}$$

In Fig. 2, the capacity bounds (17) and (18) are shown compared with the maximum user capacity of conventional MU-MIMO with linear receiver denoted as $C_{user-Conv}^{max} = m_s$ [7, 9]. As can be seen, the capacity of the proposed scheme is increased by U users of LPG. Therefore, up to *double* user capacity $(m_s + 1 \leq K \leq 2m_s)$ could be achieved which is significant in terms of serving more users for the same accessible bandwidth and available RF chains at BS.

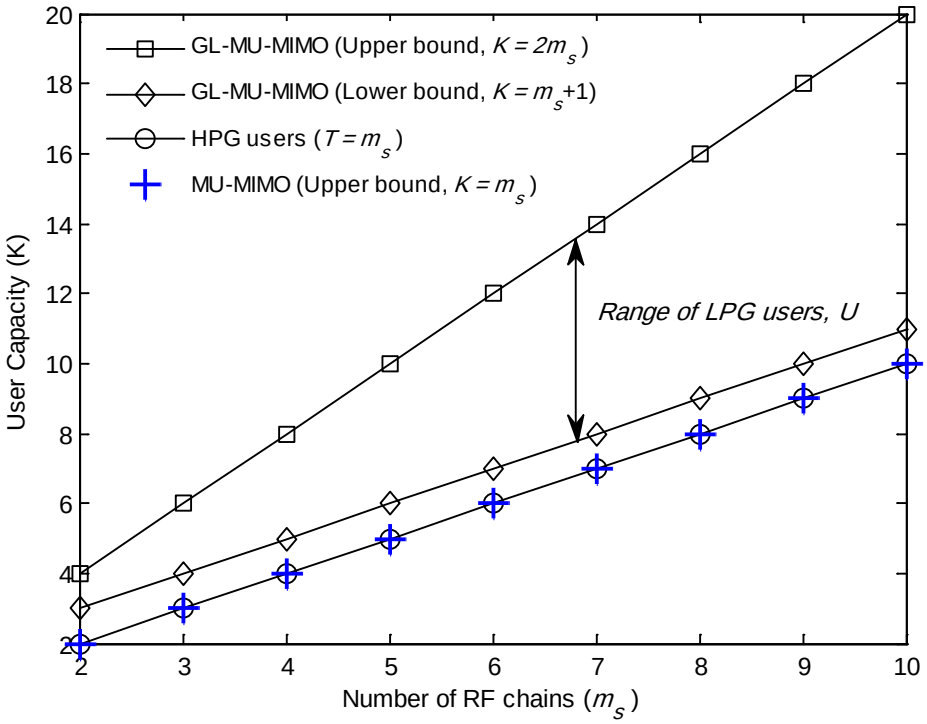

Figure 2. User capacity bounds of GL-MU-MIMO compared with the capacity of conventional MU-MIMO system with linear receiver.

3.3. Complexity Analysis

In this section, the complexity of GL-MUD is analysed and compared with ZF and MMSE receivers of conventional MU-MIMO and evaluated in terms of the required RF chains (hardware) and complex-valued multiplication efforts for signal estimation. Note that each inverse operation of a complex $N \times N$ matrix $\mathbf{A}$ requires Cholesky or Eigenvalue decomposition with cost of $N^3/6$ multiplications [31]. For fair comparison, maximum user capacity of K users is assumed for all schemes and $T = U = K/2$ for GL-MUD.

In terms of the required RF chains, at least $m_s = K/2$ is needed for GL-MUD compared to $m_s = m = K$ for ZF and MMSE receivers. The proposed GL-MUD can therefore significantly achieve up to 50% reduction in the costly hardware circuits and thus, leading to substantial reduction on consumed power and size. Accordingly, the computational efforts are calculated for GL-MUD based on the requirements of estimating $\hat{\mathbf{v}}_H$ and $\hat{\mathbf{v}}_L$ signal vectors using (8) and (9), respectively. For ZF and MMSE, transmitted vector $\mathbf{v}$ over the channel $\mathbf{G} \in \mathcal{C}^{m_s \times K}$ can be estimated from received signal $\mathbf{r}$ using $\mathbf{W}_{zf}$ and $\mathbf{W}_{mmse}$, respectively as

$$\hat{\mathbf{v}}_{zf} = \mathbf{W}_{zf}\,\mathbf{r} = \mathbf{G}^{\dagger}\,\mathbf{r} = \left[\left(\mathbf{G}^{\mathrm{H}}\mathbf{G}\right)^{-1}\mathbf{G}^{\mathrm{H}}\right]\mathbf{r} \tag{19}$$

$$\hat{\mathbf{v}}_{mmse} = \mathbf{W}_{mmse}\,\mathbf{r} = \left[\left(\mathbf{G}^{\mathrm{H}}\mathbf{G} + \sigma_n^2\mathbf{I}_K\right)^{-1}\mathbf{G}^{\mathrm{H}}\right]\mathbf{r} \tag{20}$$

Table 1 provides detailed comparison of hardware complexity and computational efforts. As can be seen, in addition to substantial low hardware complexity, the approximate essential calculations (multiplications, additions, and subtractions) of $1.04\,K^3 + 1.5\,K^2 - K$ in GL-MUD receiver is still significantly less than those for ZF and MMSE though all of them experiencing same complexity order of $\mathcal{O}(K^3)$. For example when $K = 16$, GL-MUD requires about 4627 calculations compared to 17023 and 17535 for ZF and MMSE, respectively. This is due to size reduction of $\breve{\mathbf{H}}_H \in \mathcal{C}^{m_s \times T}$ and $\breve{\mathbf{H}}_L \in \mathcal{C}^{m_s \times U}$ utilized in proposed scheme compared with $\mathbf{G} \in \mathcal{C}^{m_s \times K}$ for ZF and MMSE. Note that the combined linear ZF-SIC [13] and MMSE-SIC [15] receivers perform better than ZF and MMSE, respectively but involve higher complexity due to SIC algorithm. Hence, they are also very complex compared with proposed GL-MUD.

TABLE 1. Complexity comparison between the proposed GL-MUD and the ZF and MMSE Receivers based on hardware requirements and computational efforts. The black filled square (■) denotes already calculated terms.

Detection Scheme	Complexity				
	Hardware (RF Chains m_s)	Computational Efforts			
		Item	Computations	Total	Order
ZF	$m_s = K$	$\mathbf{G}^{\mathrm{H}}\mathbf{G}$	$K^2(2m_s - 1) = 2K^3 - K^2$	$4.16\,K^3 - K$	$\mathcal{O}(K^3)$
		$(\blacksquare)^{-1}$	$K^3/6$		
		$\blacksquare \times \mathbf{G}^{\mathrm{H}}$	$2K^2m_s - Km_s = 2K^3 - K^2$		
		$\hat{\mathbf{v}}_{zf} = \blacksquare \times \mathbf{r}$	$2Km_s - K = 2K^2 - K$		
MMSE	$m_s = K$	$\mathbf{G}^{\mathrm{H}}\mathbf{G}$	$K^2(2m_s - 1) = 2K^3 - K^2$	$4.16\,K^3 + 2K^2 - K$	$\mathcal{O}(K^3)$
		$(\blacksquare + \sigma_n^2\mathbf{I}_K)^{-1}$	$K^3/6 + 2K^2$		
		$\blacksquare \times \mathbf{G}^{\mathrm{H}}$	$2K^2m_s - Km_s = 2K^3 - K^2$		
		$\hat{\mathbf{v}}_{mmse} = \blacksquare \times \mathbf{r}$	$2Km_s - K = 2K^2 - K$		
GL-MUD	$m_s = \frac{K}{2}$	$\breve{\mathbf{H}}_H^{\mathrm{H}}\breve{\mathbf{H}}_H$	$T^2(2m_s - 1) = (K^3 - K^2)/4$	$1.04\,K^3 + 1.5\,K^2 - K$	$\mathcal{O}(K^3)$
		$(\blacksquare + \sigma_n^2\mathbf{I}_T)^{-1}$	$T^3/6 + 2T^2 = K^3/48 + K^2/2$		
		$\blacksquare \times \breve{\mathbf{H}}_H^{\mathrm{H}}$	$2T^2m_s - Tm_s = (K^3 - K^2)/4$		
		$\hat{\mathbf{v}}_H = \blacksquare \times \breve{\mathbf{r}}$	$2Tm_s - T = (K^2 - K)/2$		
		$\hat{\mathbf{s}}_H = \breve{\mathbf{H}}_H\hat{\mathbf{v}}_H$	$2Tm_s - m_s = (K^2 - K)/2$		
		$\breve{\mathbf{H}}_L^{\mathrm{H}}\breve{\mathbf{H}}_L$	$U^2(2m_s - 1) = (K^3 - K^2)/4$		
		$(\blacksquare + \sigma_n^2\mathbf{I}_U)^{-1}$	$U^3/6 + 2U^2 = K^3/48 + K^2/2$		
		$\blacksquare \times \breve{\mathbf{H}}_L^{\mathrm{H}}$	$2U^2m_s - Um_s = (K^3 - K^2)/4$		
		$\hat{\mathbf{v}}_L = \blacksquare \times (\breve{\mathbf{r}} - \hat{\mathbf{s}}_H)$	$2Um_s - U + m_s = K^2/2$		

4. Numerical Results

To demonstrate the effectiveness of proposed GL-MU-MIMO system, Monte Carlo simulations using MATLAB/v.7.9 have been carried out in this section. In MU-MIMO standards of uplink WiMAX and LTE, up to 4 users can be supported in IEEE 802.16m and 8 users in LTE Release 9-13 while Release 14 and beyond (LTE-Advanced Pro.) are still under development to meet the 5G requirements of more than 16 users towards massive MU-MIMO [1, 8]. Therefore, representative results of $4-32$ users are presented without loss of generality. For notational convenience, $K \times m/m_s$ denotes GL-MU-MIMO of K users, m receive antennas, and m_s RF chains. $K \times m$ (ZF), $K \times m$ (MMSE), and $K \times m$ (MMSE $-$ SIC) denote the reference MU-MIMO of K users and m receive antennas with ZF [3], MMSE [13], and MMSE-SIC [15] receivers, respectively. In addition, NOMA (ML) represents the NOMA for OFDM system in [32] with the optimal ML receiver. For fair comparisons, total average power is assumed to be $P = K$ for the considered schemes. To asses the error performance, we consider coherent reception of BPSK signals and results of derived BER expressions are averaged over 10^6 channel realizations while for simulations, a frame of 100 symbols is assumed for each channel realization. Furthermore, the interference cancellation error at the second layer of GL-MUD is represented in dB as $10\log_{10}\sigma_e^2$.

Error performance of HPG and LPG and the average BER for $4 \times 2/2$ GL-MU-MIMO without RAS using $\eta = 0.15,\ 0.1, 0.05$ and $\sigma_e^2 = -40$ dB are shown in Figs. 3, 4 and 5, respectively. The results are compared with the theoretical performance of 4×4 (MMSE $-$ SIC), 4×4 (MMSE), and 4×4 (ZF) that have four RF chains (i.e. two extra RF chains to serve four users). As can be seen, simulation curves are very close to that achieved by derived expressions, particularly for moderate to high SNR (> 10 dB). This demonstrate the accuracy of given analytical formulations. For HPG in Fig. 3, the error performance is directly affected by η where as η decreased, better performance is achieved due to the reduction of interference level from LPG. For example at BER of 10^{-3} and compared with theoretical results of 4×4 (MMSE $-$ SIC), significant gain of 3.6 dB is achieved using $\eta = 0.15$ with considerable reduction of essential RF chains (50%). This gain is further increased to 7 dB for $\eta = 0.1$ and 8 dB for $\eta = 0.05$. Compared with 4×4 (MMSE) and 4×4 (ZF), higher gain is achieved as expected where both of the MMSE and ZF receivers have less performance than MMSE-SIC technique. Summary of achieved gains is presented in Table 2. Note that as $\eta \to 0$, LPG interference approaches zero and HPG performance will improved considerably. On the other hand, BER of LPG in Fig. 4 depends on η and error propagation from first layer of GL-MUD. The achieved error performance of LPG has direct influence on the system average BER as shown in Fig. 5. For example, $\eta = 0.1$ provides better BER than 4×4 (ZF) by 5 dB and close to 4×4 (MMSE) by 1 dB compared with $\eta = 0.15$ and 0.05. Moreover, the controlled power difference between HPG and LPG using η will result in fairness [32] and desired unequal error protection for modern communications to provide different quality of services [33].

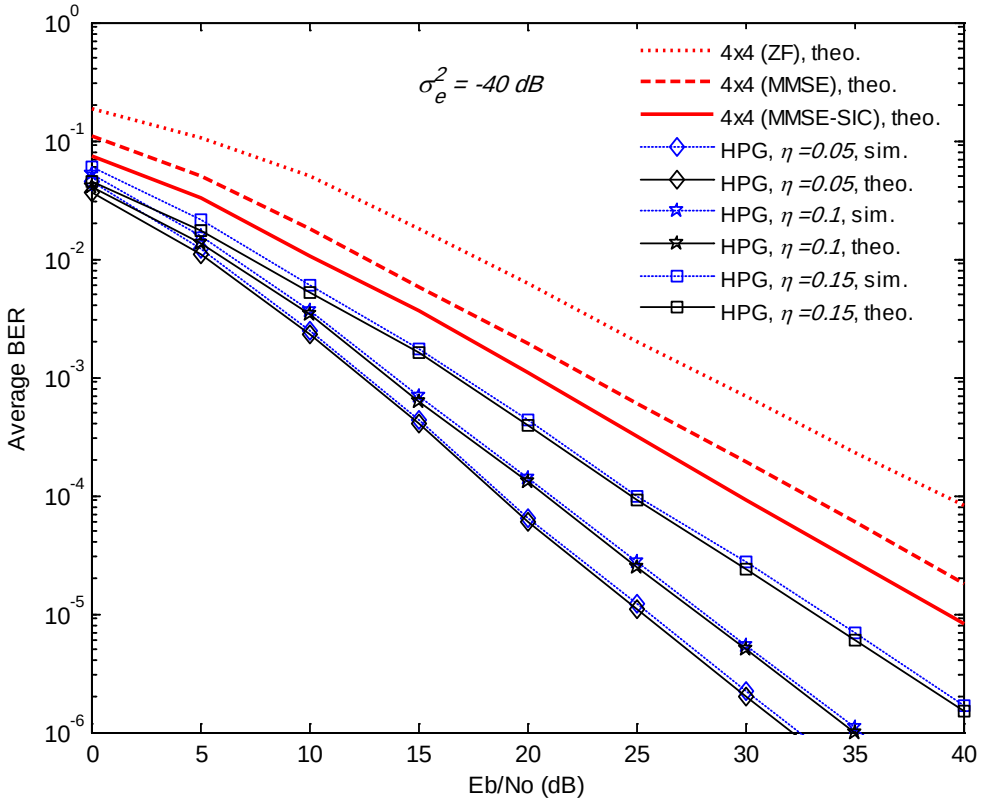

Figure 3. HPG performance of BPSK $4 \times 2/2$ system using $\eta = 0.05,\ 0.1,\ 0.15$ and $\sigma_e^2 = -40$ dB compared with 4×4 (MMSE $-$ SIC), 4×4 (MMSE), and 4×4 (ZF).

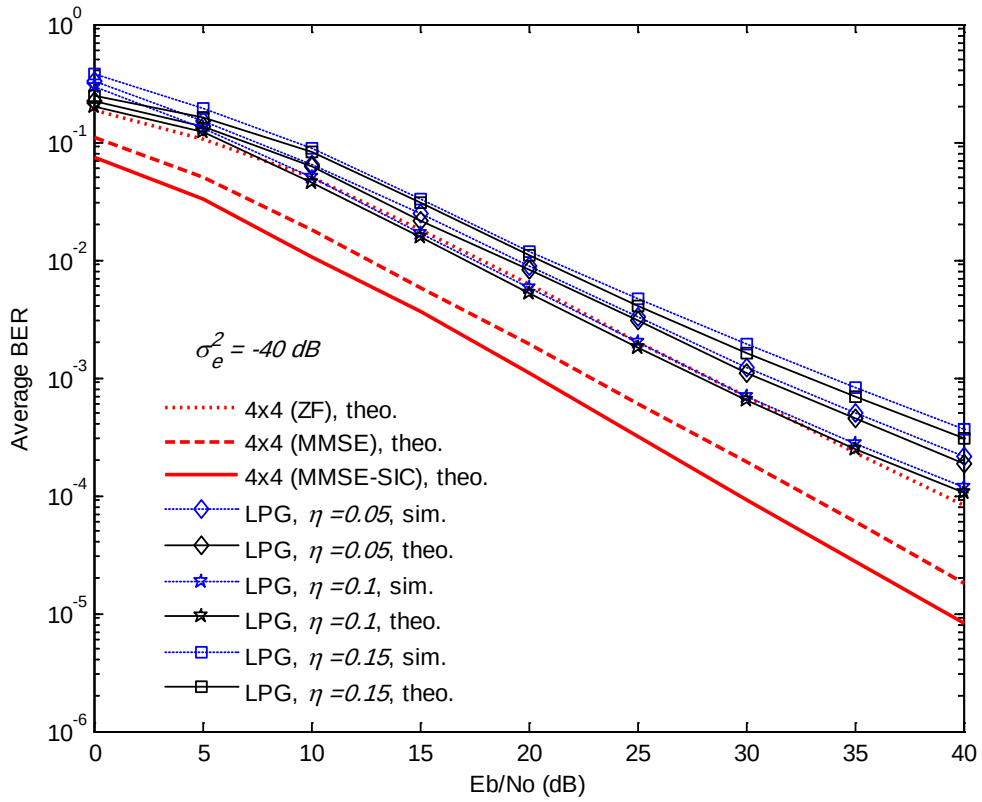

Figure 4. LPG performance of BPSK $4 \times 2/2$ system using $\eta = 0.05,\ 0.1,\ 0.15$ and $\sigma_e^2 = -40$ dB compared with 4×4 (MMSE $-$ SIC), 4×4 (MMSE), and 4×4 (ZF).

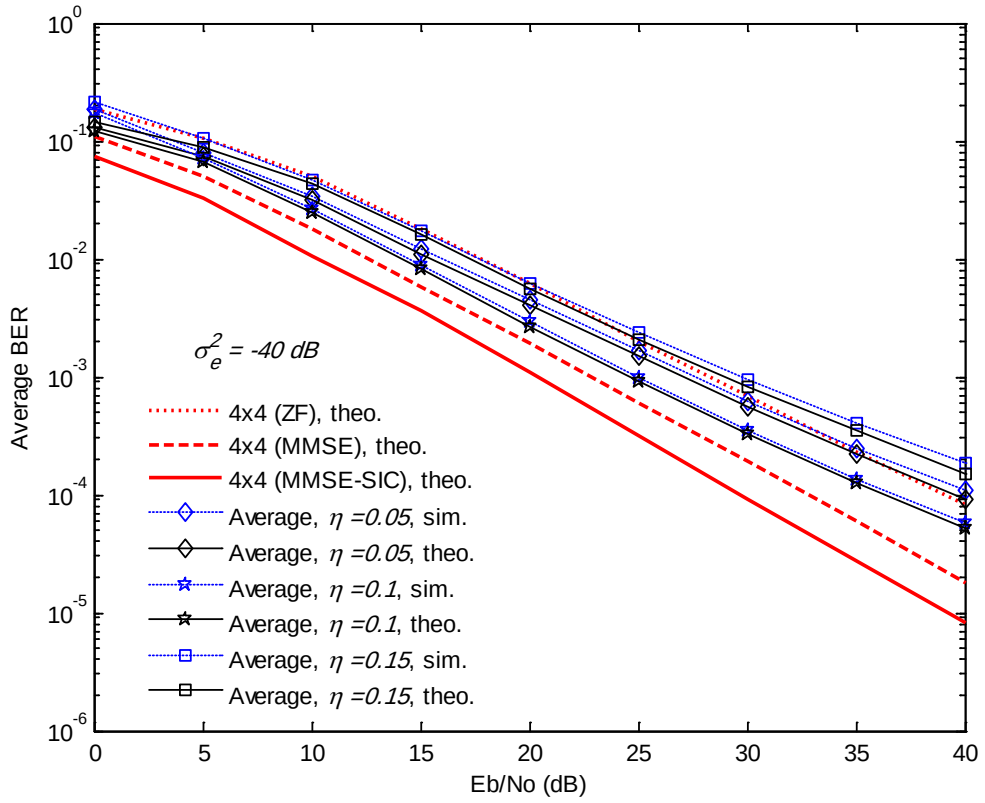

Figure 5. Average BER of BPSK $4 \times 2/2$ system using $\eta = 0.05,\ 0.1,\ 0.15$ and $\sigma_e^2 = -40$ dB compared with 4×4 (MMSE $-$ SIC), 4×4 (MMSE), and 4×4 (ZF).

TABLE 2. Summary of SNR gain of HPG in dB for BPSK $4 \times 2/2$ system using $\eta = 0.15,\ 0.1,\ 0.05$ and $\sigma_e^2 = -40$ dB compared with 4×4 (MMSE $-$ SIC), 4×4 (MMSE), and 4×4 (ZF) at BER of 10^{-3}.

MU-MIMO Scheme	SNR Gain of HPG (dB) for $4 \times 2/2$ GL-MU-MIMO		
	$\eta = 0.15$	$\eta = 0.1$	$\eta = 0.05$
4×4 (MMSE $-$ SIC)	3.6	7	8
4×4 (MMSE)	6	9	10
4×4 (ZF)	9.5	12.5	13.5

In Fig. 6, $K \times 4/4$ system with $K = 8,\ 7,\ 6,$ and 5 is examined using $\eta = 0.1,\ 0.086,\ 0.067,$ and 0.04, respectively with $\sigma_e^2 = -40$ dB to show the tradeoff between achieved user capacity gain and system error performance. Reference BER results of 4×4 (ZF), 4×4 (MMSE), 8×8 (MMSE) and 8×8 (MMSE $-$ SIC) are demonstrated also for comparison. As can be seen, theoretical and simulation outcomes become very close to each other as K decreased from 8 to 5. Moreover, at BER of 10^{-3} and considering theoretical curves, significant gain of 8 dB is achieved using $K = 5$ (i.e. HPG of $T = 4$ and LPG of $U = 1$) compared with 4×4 (ZF) and show 4.9 dB, 2.8 dB, and 1 dB gains compared with 8×8 (MMSE), 4×4 (MMSE), and 8×8 (MMSE $-$ SIC), respectively. For

$K = 6$ by adding one more user to LPG (i.e. $U = 2$), the gain of $6 \times 4/4$ is reduced to 6.3 dB, 3.4 dB, and 0.8 dB compared with 4×4 (ZF), 8×8 (MMSE), and 4×4 (MMSE), respectively. This can be explained as a tradeoff due to the increase of user capacity by two (saving of two antennas associated with RF chains). As K increased, the number of LPG users U is increased causing additional interference to the first layer of GL-MUD and hence, BER degradation. For example, $8 \times 4/4$ outperforms 4×4 (ZF) by 0.4 dB and shows less performance than 8×8 (MMSE) by 2.6 dB but at significant saving of four antennas associated with RF chains. Therefore, user capacity gain can be controlled according to required BER and through appropriate choice of η. Based on the capacity and reliability requirements, η values can be calculated and tabulated for different system configurations. Table 3 shows summary of the achieved results.

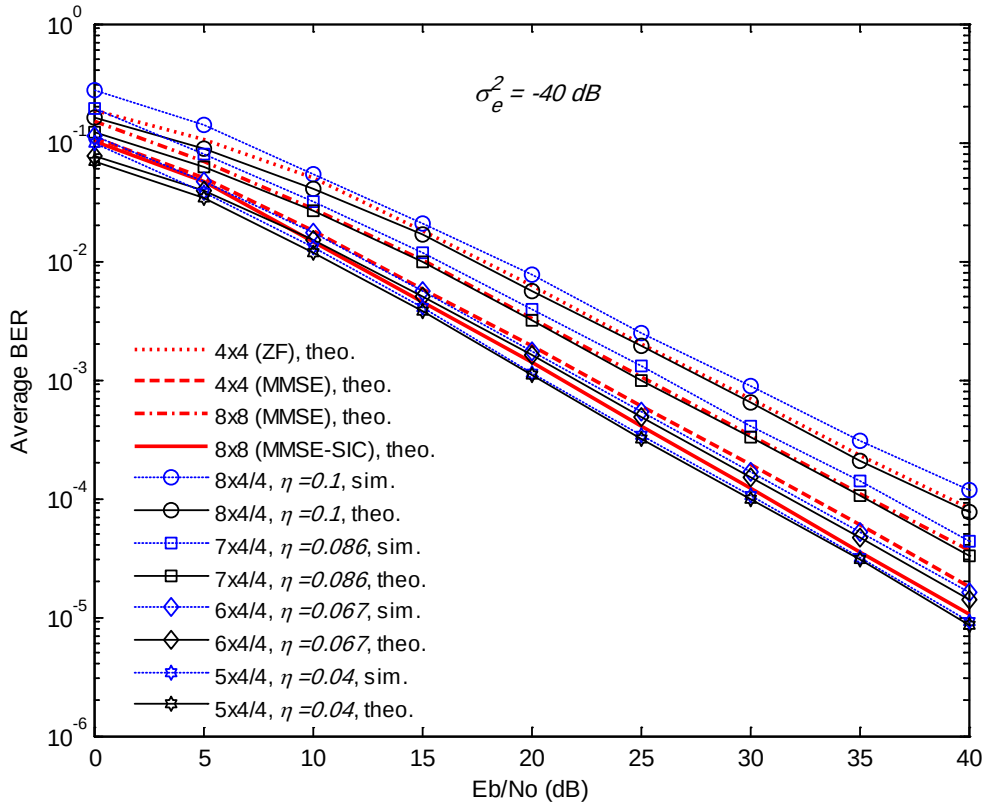

Figure 6. Average BER of BPSK $K \times 4/4$ system for $K = 8,\ 7,\ 6,$ and 5 using $\eta = 0.1,\ 0.086,\ 0.067,$ and 0.04, respectively for $\sigma_e^2 = -40$ dB compared with 4×4 (ZF), 4×4 (MMSE), 8×8 (MMSE), and 8×8 (MMSE $-$ SIC).

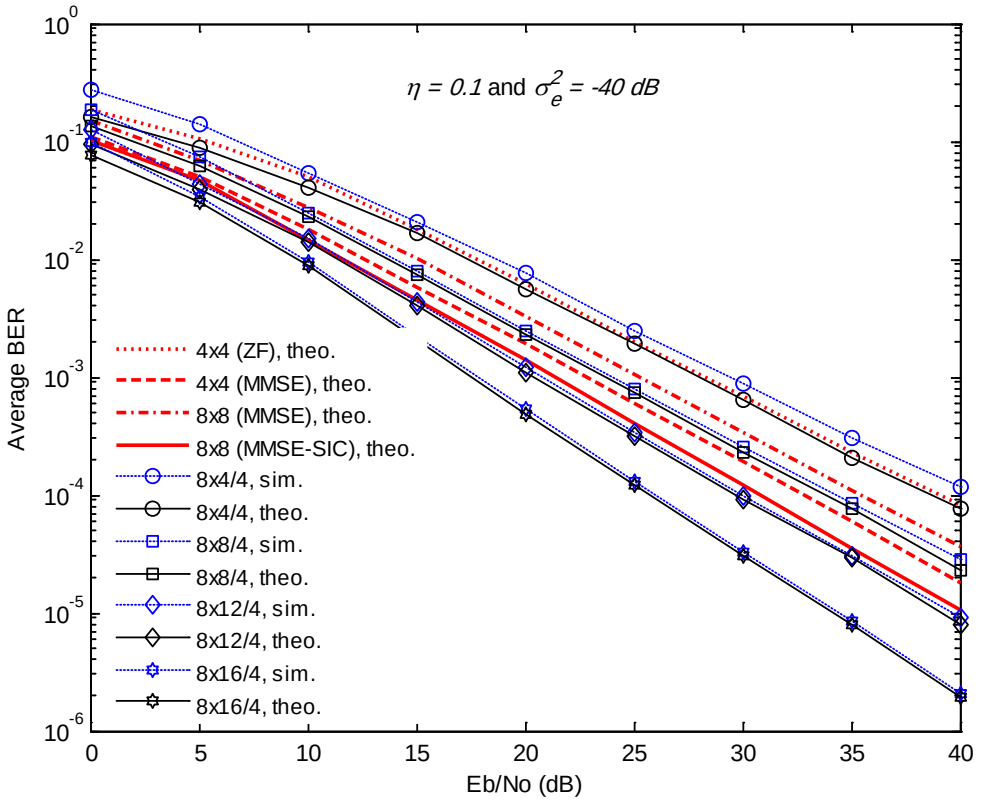

Figure 7. Average BER of BPSK $8 \times m/4$ using $\eta = 0.1$ and $m = 4,\ 8,\ 12, and\ 16$ for $\sigma_e^2 = -40$ dB compared with 4×4 (ZF), 4×4 (MMSE), 8×8 (MMSE), and 8×8 (MMSE $-$ SIC).

TABLE 3. Summary of SNR gain in dB of BPSK $K \times 4/4$ system for $K = 8,\ 7,\ 6,$ and 5 using $\eta = 0.1$, $0.086,\ 0.067,$ and 0.04, respectively for $\sigma_e^2 = -40$ dB compared with 4×4 (ZF), 4×4 (MMSE), 8×8 (MMSE), and 8×8 (MMSE $-$ SIC) at BER of 10^{-3}.

MU-MIMO Scheme	SNR Gain of GL-MU-MIMO			
	$5 \times 4/4$ $(\eta = 0.04)$	$6 \times 4/4$ $(\eta = 0.067)$	$7 \times 4/4$ $(\eta = 0.086)$	$8 \times 4/4$ $(\eta = 0.1)$
4×4 (ZF)	8	6.3	3.3	0.4
4×4 (MMSE)	2.8	0.8	-2.1	-4.9
8×8 (MMSE)	4.9	3.4	0.4	-2.6
8×8 (MMSE $-$ SIC)	1	-0.6	-3.6	-6.6

TABLE 4. Summary of SNR gain in dB of BPSK $8 \times m/4$ using $\eta = 0.1$ and $m = 4,\ 8,\ 12, and\ 16$ for $\sigma_e^2 = -40$ dB compared with 4×4 (ZF), 4×4 (MMSE), 8×8 (MMSE), and 8×8 (MMSE $-$ SIC) at BER of 10^{-3}.

MU-MIMO Scheme	SNR Gain of GL-MU-MIMO			
	$8 \times 4/4$	$8 \times 8/4$	$8 \times 12/4$	$8 \times 16/4$
4×4 (ZF)	0.4	4.5	7.7	10.7
4×4 (MMSE)	-4.9	-0.6	2.5	5.5
8×8 (MMSE)	-2.6	1.6	5	8
8×8 (MMSE $-$ SIC)	-6.6	-2	1	4

To evaluate the average BER of GL-MU-MIMO with RAS, we consider $8 \times m/4$ system configuration where $K = 8$; $m = 4, 8, 12, 16$; $m_s = 4$; $\eta = 0.1$; and $\sigma_e^2 = -40$ dB. This example demonstrate the maximum capacity of 8 users ($T = 4$ and $U = 4$) for a fixed number of RF chains ($m_s = 4$). In Fig. 7, BER results are shown compared with the theoretical outcomes of 4×4 (ZF), 4×4 (MMSE), 8×8 (MMSE), and 8×8 (MMSE $-$ SIC). As can be seen, the performance is improved considerably as m increased compared with $8 \times 4/4$ which outperforms 4×4 (ZF) by only 0.4 dB at BER of 10^{-3}. For example, the SNR gain is about 4.5, 7.7, and 10.7 dB for $m = 8$, 12, and 16 , respectively. On the other hand and compared with 8×8 (MMSE), BER of $8 \times 4/4$ which shows less performance by 2.6 dB is improved significantly as m increased. For instance, the achieved SNR gain is recorded as 1.6, 5, and 8 dB for $m = 8$, 12, and 16, respectively. Therefore, for limited m_s RF chains, RAS has direct effect to maximize the capacity and improve the BER of proposed system considerably. Summary of the achieved results is shown in Table 4. To see the impact of interference cancellation error due to HPG-MUD stage, Figs. 8 and 9 demonstrate the error performance of same system configuration using $\sigma_e^2 = -30$ dB and $\sigma_e^2 = -20$ dB, respectively. It can be seen clearly that the performance of all considered configurations is reduced as the interference error increases, and particularly for those of less RAS diversity. For example at BER of 10^{-3} and compared with the results of Fig. 7 when $\sigma_e^2 = -40$ dB, theoretical performance of $8 \times 16/4$ system is reduced by 2 dB and 13 dB for the scenarios of $\sigma_e^2 = -30$ dB and $\sigma_e^2 = -20$ dB, respectively. As the receive diversity decreases towards $m = 4$, error floor is appeared due to increased propagation error from HPG-MUD to LPG-MUD layer. Therefore, additional receive antennas should be used to mitigate the effects of interference error.

In the context of 5G requirements, GL-MU-MIMO with RAS is evaluated considering $32 \times m/16$ scenario of $K = 32$; $m = 16, 32, and\ 64$; $m_s = 16$; $\eta = 0.1$; and $\sigma_e^2 = -40$ dB. This example demonstrates maximum user capacity of 32 users ($T = 16$ and $U = 16$) for the available DoF of $m_s = 16$ RF chains. In Fig. 10, the BER outcomes are depicted compared with the results of 32×32 (MMSE $-$ SIC), 32×32 (MMSE), and NOMA (ML). As in [32], the NOMA system is simulated assuming: 64 OFDM subcarriers of 15 KHz spacing, maximum of two superimposed users per subcarrier, 10 MHz system bandwidth of 50 resource blocks, and maximum of 32 active users (26 for the generic system overloaded with 6 additional users for a given spectral efficiency target). As can be seen, the performance of proposed scheme is improved considerably as m increased. Considering the theoretical results at BER of 10^{-3} with $m = 64$, it shows substantial SNR gain of about 2.8, 4.8, and 9.6 dB compared with 32×32 (MMSE $-$ SIC), NOMA (ML), and 32×32 (MMSE), respectively. Also, it should be noted that the proposed system utilizes less number of 16 multiple-access dimensions (DoF) compared with 26 for NOMA and 32 for MMSE and MMSE-SIC systems. This has direct influence to realize the main requirements of next generation systems of high spectral efficiency, reliability, quality of service, and affordable complexity.

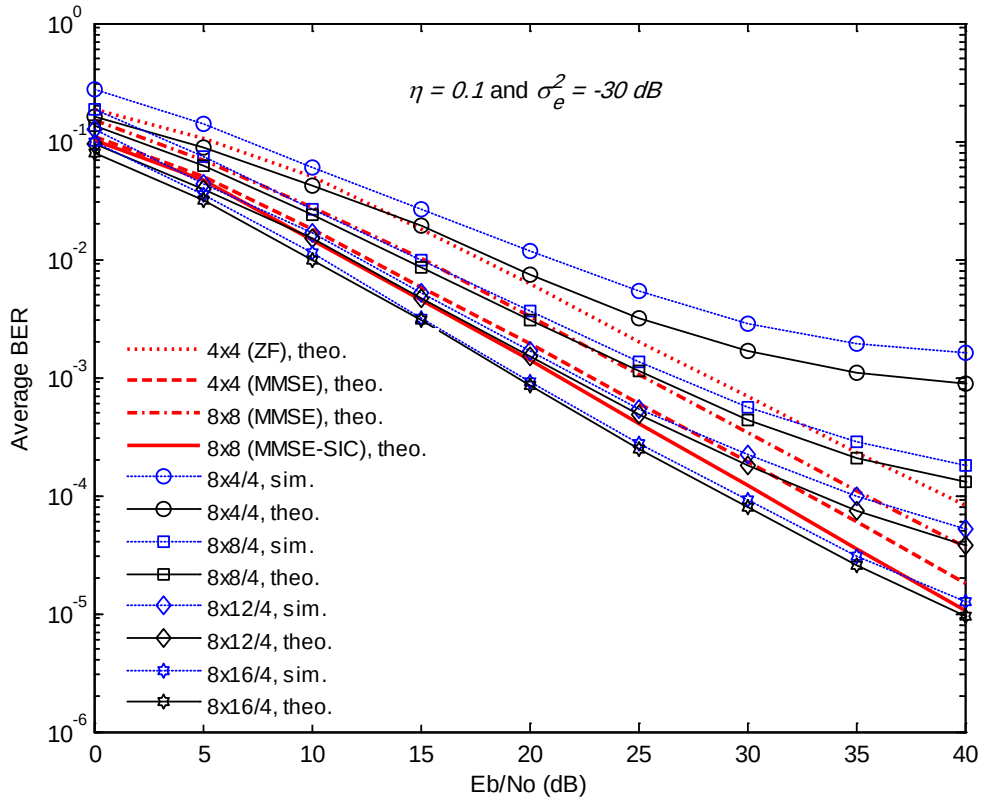

Figure 8. Average BER of BPSK $8 \times m/4$ using $\eta = 0.1$ and $m = 4, 8, 12, 16$ for $\sigma_e^2 = -30$ dB compared with 4×4 (ZF), 4×4 (MMSE), 8×8 (MMSE), and 8×8 (MMSE $-$ SIC).

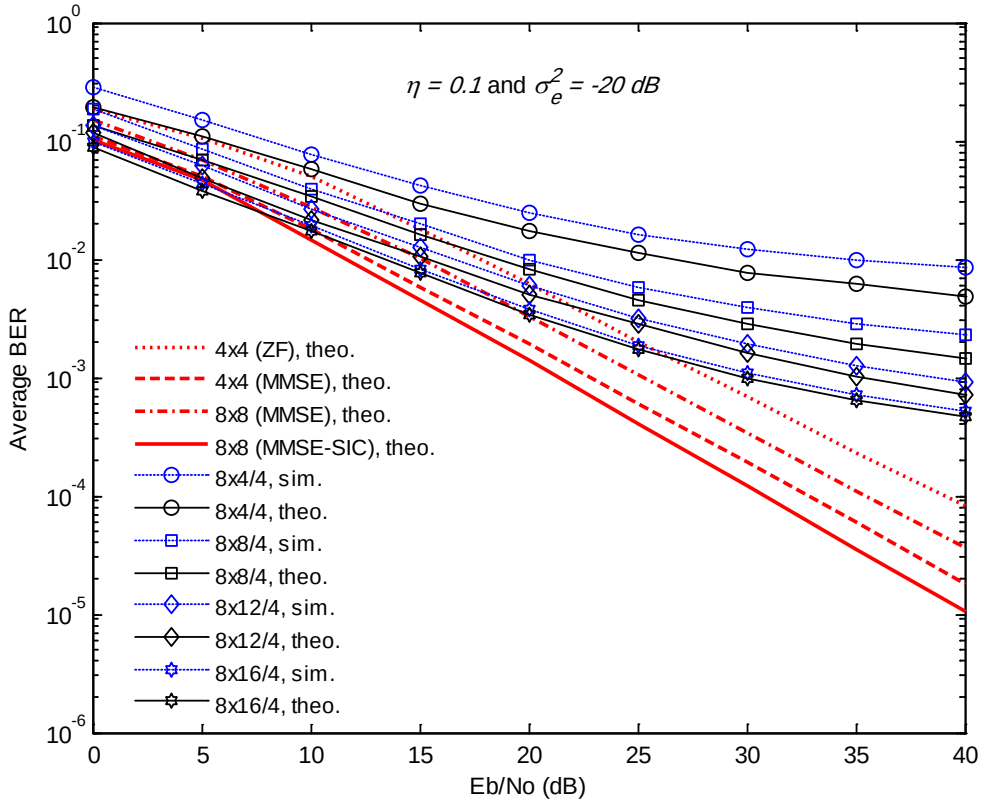

Figure 9. Average BER of BPSK $8 \times m/4$ using $\eta = 0.1$ and $m = 4, 8, 12, 16$ for $\sigma_e^2 = -20$ dB compared with 4×4 (ZF), 4×4 (MMSE), 8×8 (MMSE), and 8×8 (MMSE $-$ SIC).

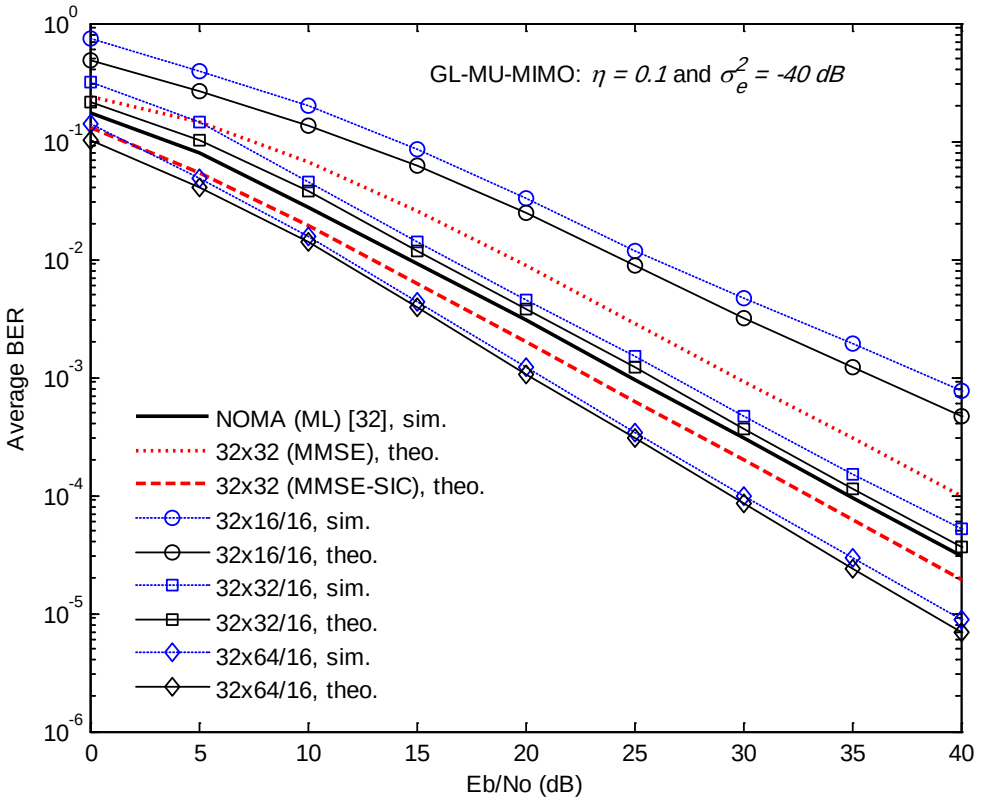

Figure 10. Average BER of BPSK $32 \times m/16$ using $\eta = 0.1$ and $m = 16,\ 32,\ 64$ for $\sigma_e^2 = -40$ dB compared with 32×32 (MMSE $-$ SIC), 32×32 (MMSE), and NOMA (ML) of 32 users (26 for the generic system overloaded by 6 users for a given spectral efficiency target).

5. Conclusions

A novel GL-MU-MIMO scheme has been proposed in this paper to extend the user capacity of conventional MU-MIMO. It has been shown to increase the capacity by up to two-fold at target BER using linear MUD approach and utilizing the same number of RF chains at BS. The proposed GL-MUD has been demonstrated to achieve significant reduction in hardware complexity (m_s RF chains) and computational efforts compared with the existing MMSE-SIC, MMSE and ZF receivers which require $2m_s$ RF chains for same number of allowed users K. The superiority of proposed scheme with GL-MUD is verified by theoretical and simulation results compared with MU-MIMO that employs linear MUD and NOMA system with ML receiver. For example, assuming user capacity of 32 and BPSK signalling at target BER of 10^{-3}, the proposed systems with $32 \times 64/16$ configuration using $\eta = 0.1$ allows 16 additional users (double user capacity for 16 RF chains) and shows significant SNR gain of 2.8 dB and 4.8 dB compared with 32×32 (MMSE $-$ SIC) that requires full number of 32 RF chains and NOMA (ML) of 26 generic and 6 extra users, respectively. The investigated scheme enables efficient spectrum utilization and reliable communications and represents a promising approach for the next 5G wireless systems.

References

[1] C. Hoymann, D. Astely, M. Stattin, G. Wikstrom, J. Cheng, A. Hoglund, M. Frenne, R. Blasco, J. Huschke, and F. Gunnarsson, "LTE release 14 outlook," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 44-49, June 2016.

[2] N. Zhao, F. R. Yu, and V. C. M. Leung, "Opportunistic Communications in Interference Alignment Networks with Wireless Power Transfer," *IEEE Wireless Commun.*, vol. 22, no. 1, pp. 88-95, Feb. 2015.

[3] A. Paulraj, D. Gore, R. Nabar, and H. Bolcskei, "An Overview of MIMO Communications-A Key to Gigabit Wireless," *Proc. IEEE*, vol. 92, no. 2, pp. 198-218, Feb. 2004.

[4] G. Miao, "Energy-Efficient Uplink Multi-user MIMO," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2302-2313, May 2013.

[5] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, *et al*., "The Role of Small Cells, Coordinated Multipoint, and Massive MIMO in 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 44 – 51, May 2014.

[6] H. Xie, B. Wang, F. Gao, and S. Jin, "A Full-Space Specrum-Sharing Strategy for Massive MIMO Cognitive Radio Systems," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 10, pp. 2537 – 2549, Oct. 2016.

[7] E. Bjornson, E. Larsson, and M. Debbah, "Massive MIMO for Maximal Spectral Efficiency: How many Users and Pilots Should be Allocated," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1293 – 1308, Feb. 2016.

[8] S. Parkvall, A. Furuskar, *et al*., "Evolution of LTE toward IMT-Advanced," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 84-91, Feb. 2011.

[9] D. Tse, P. Viswanath, and L. Zheng, "Diversity-Multiplexing Tradeoff in Multiple-Access Channels," *IEEE Trans. Inform. Theory*, vol. 50, no. 9, pp. 1859-1874, Sep. 2004.

[10] A. Soysal, and S. Ulukus, "Joint Channel Estimation and Resource Allocation for MIMO systems—PartII: Multi-User and Numerical Analysis," *IEEE Trans. Wireless Commun.*, vol. 9, no. 2, pp. 632-640, Feb. 2010.

[11] Z. Ding, F. Adachi, and H. Poor, "The Application of MIMO to Non-Orthogonal Multiple Access," *IEEE Trans. Wireless Commun*., vol. 15, no. 1, pp. 537-552, Jan. 2016.

[12] A. Zafar, M. Shaqfeh, M. Alouini, and H. Alnuweiri, "On Multiple Users Scheduling Using Superposition Coding over Rayleigh Fading Channel," *IEEE Commun. Lett.*, vol. 17, no. 4, pp. 733-736, Apr. 2013.

[13] S. Sugiura, S. Chen, and L. Hanzo, "MIMO-Aided Near-Capacity Turbo Transceivers: Taxonomy and Performance versus Complexity," *IEEE Commun. Surveys and Tutorials*, vol. 14, no. 2, pp. 421-442, 2012.

[14] X. Zhu, and R. Murch, "Performance Analysis of Maximum Likelihood Detection in a MIMO Antenna System," *IEEE Trans. Commun.*, vol. 50, no. 2, pp. 187-191, Feb. 2002.

[15] A. Zanella, M. Chiani, and M. Win, "MMSE Reception and Successive Interferance Cancellation for MIMO Systems with High Spectral Efficiency," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1244-1251, May 2005.

[16] Abhay A.K. Sah, and A. Chaturvedi, "An MMP based Approach for Detection in Large MIMO Systems using Sphere Decoding," *IEEE Wireless Commun. Letters*, DOI: 10.1109/LWC.2016.2646368, vol. PP, no. 99, pp. 1-1, Dec. 2016.

[17] B. Zarikoff, J. Cavers, and S. Bavarian, "An Iterative Groupwise Multiuser Detector for Overloaded MIMO Applications," *IEEE Trans. Wireless Commun.*, vol. 6, no. 2, pp. 443-447, Feb. 2007.

[18] W. Al-Hussaibi, and F. Ali, "Fast Receive Antenna Selection for Spatial Multiplexing MIMO over Correlated Rayleigh Fading Channels," *Wireless Personal Commun.*, vol. 70, no. 4, pp. 1243-1259, June 2013.

[19] Z. Xu, S. Sfar, and R. Blum, "Analysis of MIMO systems with receive antenna selection in spatially correlated Rayleigh fading channels," *IEEE Trans. Veh.Technol*., vol. 58, no. 1, pp. 251-262, Jan. 2009.

[20] P. Amadori, and C. Masouros, "Interference-driven antenna selection for massive multiuser MIMO," *IEEE Trans. Veh.Technol.*, vol. 65, no. 8, pp. 5944-5958, Aug. 2016.

[21] N. Mehta, S. Kashyap, and A. Molisch, "Antenna Selection in LTE: From Motivation to Specification," *IEEE Commun. Mag.*, vol. 50, no. 10, pp. 144-150, Oct. 2012.

[22] H. Haci, H. Zhu, and J. Wang, "Performance of Non-orthogonal Multiple Access (NOMA) with a Novel Asynchronous Interference Cancellation Technique," *IEEE Trans. Commun.*, vol. PP, no. 99, pp. 1-1, DOI: 10.1109/TCOMM.2016.2640307, Jan. 2017.

[23] H. Sari, F. Vanhaverbeke, and M. Moeneclaey, "Extending the capacity of multiple access channels," *IEEE Commun. Mag.*, vol. 38, pp. 74 – 82, Jan. 2000.

[24] S. Bopping and J.M. Shea, "Superposition Coding in the Downlink of CDMA Celluar Systems," *in Proc. IEEE WCNC'06*, vol. 4, pp. 1978-1983, 2006.

[25] L.L. Yang, "MIMO-Assisted Space-Code-Division Multiple-Access: Linear Detectors and Performance Over Multipath Fading Channels," *IEEE J. Sel. Areas Commun*., vol. 24, no. 1, pp. 121-131, Jan. 2006.

[26] F. Ali, and I. Shakya, "Collaborative Spreading for the Downlink of Overloaded CDMA," *Wireless Commun. Mob. Comput.*, vol. 10, no. 3, pp. 383 – 393, Mar. 2010.

[27] S. Peters, and R. Heath, Jr., "User Partitioning for Less Overhead in MIMO Interference Channels," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 592-603, Feb. 2012.

[28] D. Tse, and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge Univ. Press, 2005.

[29] J. G. Proakis, *Digital Communications*, 4th edition, New York: McGraw-Hill, 2001.

[30] H. Keshavarz, L.-L. Xie, and R. Mazumdar, "User Capacity Scaling Laws for Fading Multiple-Access Channels," *IEEE Trans. Wireless Commun.*, vol. 8, no. 9, pp. 4498-4507, Sep. 2009.

[31] W. Al-Hussaibi, and F. Ali , "Generation of Correlated rayleigh fading Channels for Accurate Simulation of promissing Wireless Communication Systems," *Simulation Modelling Practice and Theory*, vol. 25, no. 4, pp. 56-72, 2012.

[32] M. Al-Imari, P.Xiao, M. A. Imran, and R. Tafazolli, "Uplink Non-Orthogonal Multiple Access for 5G Wireless Networks," *in Proc. 11th IEEE ISWCS*, pp. 781-785, 26-29 Aug. 2014.

[33] T. Tillo, E. Baccaglini, and G. Olmo, "Unequal Protection of Video Data According to Slice Relevance," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1572-1582, June 2011.